# EXHIBIT M

# Glossary



These concepts are not unique to Anthropic's language models, but we present a brief summary below:

## Context Window

The "context window" refers to how much text a language model can look back on and reference, when attempting to generate text. This is different from the large corpus of data the language model the was trained on, and instead represents more of a "working memory" for the model. With Claude Slackbot, the context window contains everything in the individual slack thread — see our section on prompt length for the current length.

## Fine-Tuning

Fine-tuning refers to the process of using additional data to further train a pretrained language model. This causes the model to start representing and mimicking the fine-tuning dataset. Claude is not a bare language model; it has already been fine-tuned to be a helpful assistant. Our API does not offer fine-tuning, but please ask your Anthropic contact if you are interested.

## HHH

These three H's represent Anthropic's goals in ensuring that Claude is beneficial to society.

- A helpful AI will attempt to perform the task or answer the question posed.
- An honest AI will give accurate information, and not hallucinate or confabulate.
- A harmless AI will not be offensive or discriminatory, and when asked to aid in a dangerous act, the AI should politely refuse.

## LLM

Large Language Models (LLMs) are AI language models with many parameters that are able to perform a variety of surprisingly useful tasks. Claude is a conversational assistant, based on a large language model.

## Pretraining

Pretraining refers to the process of training language models on a large unlabeled corpus of text. In Claude's case, autoregressive language models (like Claude's underlying model) are pretrained to predict the next word, given the previous context of text in the document. These models are not good at answering questions or following instructions, and often require a deep skill in prompt engineering to elicit behaviors. It's through fine-tuning and RLHF that these pretrained models become useful for many tasks.

## RLHF

Reinforcement Learning from Human Feedback is a means to take a pretrained language model, and encourage it to behave in ways that are consistent with with humans prefer. This can include "helping it to follow instructions" or "helping it to act more like a chat bot". The human feedback consists of a human-ranking set of two or more examples text, and the reinforcement learning encourages the model learns to prefer outputs that are similar to the higher-ranked ones. Claude is not a bare language model; it has already been trained with RLHF to be a helpful assistant.. For more details, you can read Anthropic's paper on the subject.

## Temperature

Temperature is a parameter that controls the randomness of a model's predictions during generation. Higher temperature leads to more creative samples that enable multiple variations in phrasing (and in the case of fiction, variation in answers as well), while lower temperature leads to more conservative samples that stick to the most-probable phrasing and answer. Adjusting the temperature is a way to encourage a language model to explore rare, uncommon, or surprising next words or sequences, rather than only selecting the most likely predictions. Claude Slackbot uses a non-zero temperature when generating responses, that allow some variation in its answers.

## Tokens

Tokens are the smallest individual "atoms" of a language model, and can varyingly correspond to words, subwords, characters, or even bytes (in the case of Unicode). For Claude the average token is about 3.5 characters. Tokens are typically hidden when interacting with language models at the "text" level, but become relevant when digging into the exact inputs and outputs of a language model. When Claude is provided language to evaluate, the language text (consisting of a series of characters) is encoded into a series of tokens for the model to act on. Larger tokens enable data-efficiency at inference time and pretraining (and so are utilized when possible), while smaller tokens enable a model to handle uncommon or never-before-seen words.

Updated 5 months ago

← Claude for Google Sheets

Introduction to prompt design →